Development of a Change Model for a Controlled Medical Vocabulary

Diane E. Oliver, M.D.[1,2] and Yuval Shahar, M.D., Ph.D.[2]

[1]VA Palo Alto Health Care System
[2]Section on Medical Informatics, Stanford University
oliver@smi.stanford.edu, shahar@smi.stanford.edu

ABSTRACT

Managing change in controlled medical vocabularies is labor intensive and costly, but change is inevitable if vocabularies are to be kept up to date. The changes that are appropriate for a controlled medical vocabulary depend on the data stored for that vocabulary, and those data in turn depend on the needs of users. The set of change operations is the **change model**; *the data stored about concepts comprise the* **concept model**. *Because the change model depends directly on the concept model, a discussion of the former necessitates a discussion of the latter. In this paper, we first present a set of tasks that we believe controlled medical vocabularies should handle. Next, we describe our concept model for a controlled medical vocabulary. Then, we review the literature on changes in existing vocabulary systems. Finally, we present our change model. We call our system, which incorporates the concept model and change model, the General Online Dictionary of Medicine (GOLDMINE).*

INTRODUCTION

Maintainers of controlled medical vocabularies recognize that such vocabularies are not static [1, 2]. Change is inevitable because medical care and medical terminology are constantly evolving. A *change model* is a specification of the changes that must be supported in a dynamic controlled medical vocabulary. The change model depends on a particular *concept model*, which specifies information stored about concepts and constraints.

If a controlled medical vocabulary is shared by different sites that deliver health care, vocabulary managers at those sites may want to modify the vocabulary locally to suit their own needs. However, they also may want to incorporate updates from the shared vocabulary. We envision that such updates will be more manageable if both the local and the shared versions of the vocabulary conform to the same concept model and change model.

To develop a change model, we have drawn from previous work in two areas: controlled medical vocabularies and frame-based knowledge-representation systems.

A number of well-known controlled medical vocabularies exist, such as the International Classification of Diseases, Ninth Edition, Clinical Modification (ICD-9-CM) [3]; Medical Subject Headings (MeSH) [4]; the Systematized Nomenclature of Human and Veterinary Medicine (SNOMED) [5]; the Diagnostic and Statistic Manual, Fourth Edition (DSM-IV) [6]; the Unified Medical Language System (UMLS) [7]; and the Medical Entities Dictionary (MED) [1]. Each vocabulary was created for a different purpose, but all must keep up with changes in medical content.

Knowledge-representation researchers have also been interested in the description and organization of concepts. Their emphasis has been on developing formal methods of representation, rather than on developing content. KL-ONE [8] was an early frame-based knowledge-representation system, created in the 1970s, that emphasized IS-A hierarchies of concepts, relationships among those concepts, and automatic classification. CLASSIC [9] was one of several frame-based knowledge-representation systems that was influenced strongly by KL-ONE. More recently, Karp developed a protocol, called the Generic Frame Protocol (GFP) [10], that specifies allowable queries and changes to frame-based knowledge-representation systems in general.

Our goal was to develop a change model that supports a particular concept model that supports tasks that are essential for a controlled-medical-vocabulary system. The system that incorporates our models is called the *General Online Dictionary of Medicine (GOLDMINE).*

METHODS

First, we identified a set of tasks that we believe are important for controlled medical vocabularies designed for clinical purposes. We based our choices on the experiences and ideas of other researchers [1, 2, 7, 8, 9, 10, 11, 12, 14] and on our own assessment.

Second, we developed the concept model for GOLDMINE, selecting features and constraints that support the specified tasks.

1091—8280/97/$5.00

Third, we reviewed the literature on changes that occur in existing controlled medical vocabularies (ICD-9-CM, MeSH, SNOMED, DSM-IV, UMLS, and MED) and frame-based knowledge-representation systems (CLASSIC and Generic Frame Protocol). We looked at the types and frequencies of changes that were made to particular versions of controlled medical vocabularies For CLASSIC and GFP, we looked at the change operations permitted.

Finally, based on the GOLDMINE concept model and our assessment of advantages and disadvantages of existing systems, we developed the change model for GOLDMINE.

TASKS FOR MEDICAL VOCABULARY SYSTEMS

A clinical information system that integrates patient databases and applications for data entry and retrieval, clinical decision support, and aggregation of patient data requires a common representation of clinical data elements. In such an environment, the meaning of a data element must remain constant over time, even if commonly used terminology changes; users must be able to browse and search for terms in the vocabulary; and applications that interact with the database must be able to use terms that are more general or more specific than the concepts stored for a particular patient. We believe that the system should also serve as both a dictionary and a thesaurus to enable users to use concepts consistently.

We have identified 10 essential tasks:

1. Given a concept name, determine the coded unique identifier for that concept, where the meaning of the identifier never changes (also called *coding*).
2. Given a coded unique identifier, determine the name of the concept to which that identifier refers (also called *decoding*).
3. Given an input string that represents a concept that the user has in mind, find a set of possible concepts that may have the same meaning as the desired concept.
4. Given a concept name or code, retrieve the text definition, synonyms, acronyms, or other abbreviations to obtain additional information on the intended meaning of that concept.
5. Given a synonym, an acronym, or other abbreviation, retrieve the concepts to which it refers.
6. Given a concept name or code, retrieve the concepts that are more general or more specific than the concept specified.
7. Determine whether a concept is more general or more specific than another concept.
8. Search for a concept by navigating a hierarchical display of concepts.
9. Determine the characteristics that make a concept similar to or different from its child, parent, ancestor, descendant, or sibling in the generalization–specialization hierarchy.
10. Translate a given concept into a standardized vocabulary, such as ICD-9-CM, MeSH, or SNOMED.

THE GOLDMINE CONCEPT MODEL

We have created the GOLDMINE concept model to address the 10 essential tasks. The GOLDMINE concept model specifies the representation of concepts and their defining attributes. There are eight concept features:

1. *Concept unique identifer*: Unique code for the concept that never changes
2. *Concept name*: Unique name for the concept that may change
3. *Concept definition*: Natural-language definition for the concept
4. *Synonyms*: Terms that can be used interchangeably in some context with the concept name
5. *Acronyms and other abbreviations*: Shortened forms of the concept name or synonyms
6. *UMLS code*: UMLS code that represents the same concept
7. *IS-A parents*: More generalized concepts (multiple parents allowed)
8. *Defining-attribute set*: Attribute–value pairs that define the concept

There are three attribute features:

1. *Attribute unique identifier*: Unique code for the attribute that never changes
2. *Attribute name*: Unique name for the attribute that may change
3. *Attribute definition*. Natural-language definition for the attribute

Attributes currently are not stored in a hierarchy in GOLDMINE. A sample concept is shown in Figure 1.

Figure 1. Sample concept in GOLDMINE. Adapted from MeSH and UMLS.

Concept unique identifier: 1000
Concept name: Ebola hemorrhagic fever
Concept text definition: a highly fatal hemorrhagic fever, clinically very similar to Marburg virus disease, caused by the Ebola virus, first occurring in the Sudan and adjacent northwestern Zaire
Synonyms: Ebola virus disease
Acronyms / Abbreviations:
UMLS code: C0282687
IS-A parents: viral hemorrhagic fever
Defining-attribute set:
caused-by: Ebola virus

A unique identifier that remains constant over time should have no inherent meaning of its own, because a term used for a concept today may not be appropriate tomorrow. For example, if vocabulary maintainers in the past had used the term *lues* as a unique identifier that we could never change, the term would be confusing to those of us who are now more familiar with the term *syphilis* and who do not recognize or even know how to pronounce the term *lues*. Therefore, we believe that a code should be used to provide constancy over time, and that a separate unique name should also be part of the model to ensure that the name can reflect common usage.

An important use of synonyms, acronyms, and other abbreviations is assisting a user to search for a concept. Substring matching of an input string with concept name, synonyms, and acronyms is likely to yield improved results (i.e., greater recall) than is matching with only the concept name.

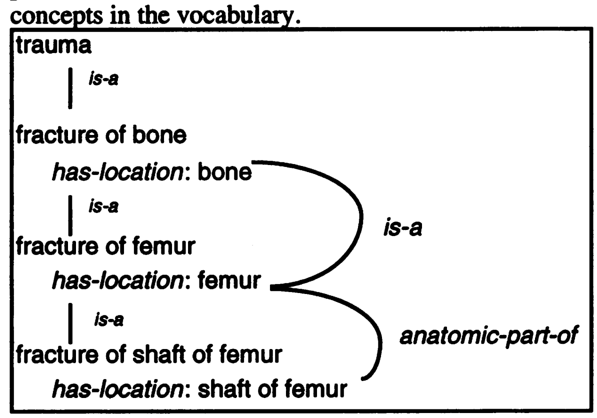
One of the tasks supported by our model is translation to other standardized vocabularies by references to the UMLS in GOLDMINE. Each GOLDMINE concept can have a UMLS code asigned to it. This code provides a translation to standardized vocabularies contained in the UMLS. Such an approach was proposed by Cimino and colleagues, who implemented it in the MED [12].

An IS-A hierarchy allows retrieval of the more general and more specific concepts related to a given concept in the vocabulary, or alternatively, determination of whether one particular concept is more general or more specific than another particular concept.

The defining-attribute set in GOLDMINE is a set of attribute–value pairs that describe the concept, including attribute–value pairs that a concept shares with its ancestors in the hierarchy, as well as those that differentiate a concept from its parents. Unfortunately, it is often difficult to specify defining attributes, or necessary and sufficient conditions, for medical concepts. Brown and colleagues found that 30 percent of medical disorders could not be defined in this way [13]. An IS-A relationship between two concepts implies certain constraints on the defining-attribute set of the child, given the defining-attribute set of the parent; however, attributes are not required for an IS-A relationship to exist between two concepts.

GOLDMINE specifies constraints on the attribute–value pairs that differentiate a child concept from the associated parent concept. For example, as shown in Figure 2, the concept *trauma* has no attribute–value pairs; its child, *fracture of bone*, adds the attribute–value pair *has-location: bone*. A child of *fracture of bone*, *fracture of femur*, refines the value of the attribute *has-location*. In this case, the child's value, *femur*, is related to the parent's value, *bone*, by an *is-a* relationship. The next child, *fracture of shaft of femur*, refines the value of *has-location* with the value *shaft of femur*. In this case, the child's value, *shaft of femur*, is related to the parent's value, *femur*, by an *anatomic-part-of* relationship. Bernauer describes such subsumptive relationships [14].

Figure 2. Sample IS-A hierarchy, showing allowable relationships between a child's attribute value and its parent's attribute value. All attribute values are concepts in the vocabulary.



CHANGES IN EXISTING VOCABULARY SYSTEMS

Changes made to existing vocabularies are reported in various ways. For example, the metaterminology, or terminology about concepts, varies. In addition, the availability of data about changes is variable among systems, and there are no standards for reporting changes. We review briefly changes in selected systems.

ICD-9-CM Changes. ICD-9-CM [3] has *codes* that identify *categories*, which are comparable to concepts; *descriptions*, which are similar to, but less detailed than, text definitions; and *exclusions*, which have no counterpart in GOLDMINE. Between 1993 and 1994, of more than 17,000 codes, 120 codes were added, 27 codes were deleted, and 16 codes were revised [3]. When a category is deleted, information about it is no longer kept in the vocabulary. A deleted code can be reused at a later time for a category with a different meaning. Descriptions and exclusions can be added, deleted, or revised.

MeSH Changes. MeSH [4] has *headings*, which are like concepts; *print entry terms*, which include synonyms and near synonyms; *tree numbers*, which specify location in a hierarchy; and *scope notes*, which serve as text definitions. In the 1997 edition of MeSH [4], of a total of 17,895 headings, 350 headings were added, 60 headings were deleted, 71 heading names were changed, and 560 print entry terms were added. There were also 687 additions and 508 deletions of tree numbers. Scope notes also can be changed.

SNOMED Changes. SNOMED [5] has *term codes*, which uniquely identify concepts; a *primary* or *preferred term* for each term code; *synonyms*, which can be identified by their term code being the same as that of the preferred term; and *references*, which are unnamed links to other concepts in the vocabulary. In SNOMED Version 3.1, there were 137,164 terms, of which 4,523 were new [5]. Other changes included deletions of terms (preferred terms or synonyms); text corrections of terms; additions or corrections of references; additions or corrections of linked ICD-9-CM codes; and reassignments of term codes. A term-code reassignment is analogous to a change in a concept's parents and children.

DSM-IV Changes. Like ICD-9-CM, DSM-IV [6] has *categories*. Changes that occurred in DSM-IV included adding a category, deleting a category, renaming a category, merging two concepts into a single category, subsuming a category by another existing category, splitting one category into multiple categories, moving a category to another grouping, and changing the text description.

UMLS Changes. The UMLS Metathesaurus publishes a set of differences between a new release and the previous version. From this set, a user can determine what concepts have been added and what changes have been made to existing concepts [15]. For concepts that are no longer in the Metathesaurus, there are two possibilities: (1) a concept could simply have been deleted, or (2) it could have been merged with another concept. To indicate whether each change was a merge or a deletion, the UMLS developers release a file of merges and a file of deletions.

MED Changes. The MED incorporates all of ICD-9-CM. Because ICD-9-CM allows reuse of codes for different concept meanings and the MED does not, this difference in concept-modeling philosophy causes problems in updating the MED [1]. Another difference is that the MED *retires* obsolete concepts by flagging them as retired, whereas ICD-9-CM deletes them.

CLASSIC Change Operations. A CLASSIC knowledge base contains concepts and individuals [8]. *Concepts* are organized into a hierarchy, and *individuals* are instances of those concepts. Controlled medical vocabularies typically do not contain both concepts and individuals; instead, they contain only concepts, and a patient database contains individuals. Therefore, in analyzing change operations in CLASSIC, we consider only changes to concepts, rather than changes to individuals.

A formal concept definition in CLASSIC specifies parents, roles, and role restrictions. A *role* is a relation. An example of a *role restriction* is the *role value type*, or concept, to which the value of that role is restricted. If the concept definition provides necessary and sufficient conditions for recognizing a concept, then the concept can be classified automatically. If it does not, then the concept is a *primitive concept*, and it cannot be classified automatically.

CLASSIC supports multiple names for a concept. Every concept must have one *canonical concept name* and may have one or more *synonyms*. The user can change the canonical concept name or add or remove synonyms.

In CLASSIC, once a concept has been defined, its definition cannot be modified, and the concept cannot be deleted. Therefore, it is not possible to change a concept's parents, roles, or role value types without creating a new concept. Such an approach is not desirable for a vocabulary that must allow uniquely identified concepts to change location in the hierarchy.

GFP Change Operations. GFP [10] has *frames*, which are either classes or instances. *Class* is used synonymously with *concept*, and *instance* is used synonymously with *individual*. A *slot* in GFP is analogous to a *role* in CLASSIC. *Superclasses* are analogous to parents.

In GFP, the maintainer can make changes to classes by using the operations *create frame (or create class)*, *rename frame*, or *delete frame*. It is also possible to make changes to slots by using *create slot*, *delete slot*, or *rename slot*. The maintainer can make changes to parents with *put class supers*, or make changes to slot values with *put slot values*, *put slot value*, *add slot value*, *replace slot value*, or *remove slot value*. GFP does not support synonyms or text definitions directly.

THE GOLDMINE CHANGE MODEL

Given the GOLDMINE concept model and an understanding of changes in existing vocabulary systems, we have designed the GOLDMINE change model. We divide the set of change operations into those that do not affect the hierarchy and those that do.

Change operations that do not affect the hierarchy are the following:

1. Replace concept name
2. Replace concept definition
3. Add synonym
4. Delete synonym
5. Add acronym/abbreviation
6. Delete acronym/abbreviation
7. Replace UMLS code
8. Add attribute
9. Replace attribute name
10. Replace attribute definition

Changes that do not affect the hierarchy have only a few simple constraints. For example, for the operation *replace concept name*, the system must make sure that the new concept name has not already been used for another concept.

GOLDMINE change operations that do affect the hierarchy are the following:

1. Add concept
2. Retire concept
3. Add IS-A parent
4. Remove IS-A parent
5. Add attribute-value pair
6. Remove attribute-value pair
7. Replace attribute value
8. Merge two or more concepts into one concept
9. Split one concept into two or more concepts

Changes that do affect the hierarchy require checks to be sure that the integrity of the hierarchy remains intact. Constraints in the GOLDMINE concept model—such as that a child concept may add attribute-value pairs to those of its parents, but may not remove any, or that a child may refine the value of a parent's attribute, but may not override that parent's attribute value—are constraints that must not be violated by changes.

SUMMARY AND FUTURE WORK

The GOLDMINE change model is specific to the concept model that underlies it. This change model is faithful to the types of changes that actually occur in existing controlled medical vocabularies, and it adopts certain features of frame-based knowledge-representation systems. However, the GOLDMINE change model differs from those of frame-based knowledge-representation systems because it does not share the concept model that distinguishes between concepts and individuals (or classes and instances), and GOLDMINE does not have automatic classification.

The tasks that we have specified for controlled medical vocabularies are not the only ones that are possible. If the task list were augmented, then the concept model may need to be augmented, and such a change would also alter the change model.

We currently are implementing a system according to the GOLDMINE concept model and change model. We are investigating methods of managing local divergence of a shared vocabulary when both the local and shared versions conform to the same models.

ACKNOWLEDGEMENTS

Dr. Oliver is supported by the VA medical-informatics training program. Dr. Shahar is supported by NLM grants LM05708 and LM06245 and NSF grant IRI-9528444. Computing resources were provided by the CAMIS project, funded under NLM grant LM05305. Additional support was provided by NLM contract N01-LM-4-3514. We are grateful to Lyn Dupre for editing this manuscript.

REFERENCES

1. Cimino JJ. Formal descriptions and adaptive mechanisms for changes in controlled medical vocabularies. *Methods of Information in Medicine* 1996; 35: 202–210.

2. O'Neill M, Payne C, Read J. Read Codes Version 3: A user led terminology. *Methods of Information in Medicine* 1995; 34: 187-192.

3. *International Classification of Diseases, Ninth Revision, Clinical Modification,* Fourth Edition. Practice Managment Information Corportation, Los Angeles, CA, 1994.

4. *MeSH: Annotated Alphabetic List*. National Library of Medicine, Bethesda, MD, 1997.

5. SNOMED International. *Addendum to Introduction to SNOMED*, Version 3.1. College of American Pathologists, Northfield IL, 1995.

6. Frances A, Pincus HA, First MB (Ed.). *Diagnostic and Statistical Manual*, Fourth Edition. American Psychiatric Association, Washington, D.C., 1994.

7. *UMLS Knowledge Sources*, Eighth Edition Documentation, National Library of Medicine, Bethesda, MD, 1997.

8. Brachman RJ, Schmolze JG. An overview of the KL-ONE knowledge-representation system. *Cognitive Science* 1985; 9: 171–216.

9. Resnick LA, Borgida A, Brachman RJ, McGuinness DL, Patel-Schneider PF, Zalondek KC. *CLASSIC Description and Reference Manual*, Version 2.2, 1993.

10. Karp PD, Gruber T. The generic frame protocol. http://www.ai.sri.com/~gfp/spec/paper/paper.html. March 29, 1997.

11. Rector AL, Bechhofer S, Goble CA, Horrocks, Nowlan WA, Solomon WD. The GRAIL concept modelling language for medical terminology. *Artificial Intelligence in Medicine* 1997; 9: 139-171.

12. Cimino JJ, Johnson SB, Hripcsak G, Sideli RV, Fink DJ, Friedman C, Clayton PD. One year's experience with the UMLS in academia and patient care. *Medinfo* 1992; 7(Pt 2): 1501–1505.

13. Brown P, personal communication, Jan. 1997.

14. Bernauer J. Subsumption principles underlying medical concept systems and their formal reconstruction. In Ozbolt JG (ed.), *Proceedings of the Eighteenth Annual SCAMC*, Washington, D.C., Hanley and Belfus, 1994; 140–144.

15. Olson NE, Erlbaum MS, Tuttle MS, Sherertz DD, Suarez-Munist O, Lipow SS, Cole WG, Nelson SJ. Exploiting the Metathesaurus update model. In Cimino JJ (ed.) *Proceedings of the Annual AMIA Fall Symposium*, Washington, D.C., Hanley and Belfus, 1996; 902.